POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

A Novel Hybrid Scoring Function for Extreme-Scale Virtual Screening in Drug Discovery

Laurea Magistrale in Computer Science and Engineering - Ingegneria Informatica

Author: ZHANG YUEDONG

Advisor: Prof. Gianluca Palermo

Co-advisor: Davide Gadioli, Gianmarco Accordi

Academic year: 2022-23

1. Introduction

Traditionally, drug discovery has been challenged by lengthy development cycles, high costs, and significant failure rates. The integration of various in-silico methods has alleviated these issues. In-Silico Drug Discovery comprises two primary stages: Target Discovery and Lead Discovery. The first stage involves accurately identifying biological macromolecules representing the disease. The second stage focuses on finding a set of small molecules that represent potential drug candidates within a large-scale molecular database for subsequent wet laboratory experiments. The leads screened from this stage are expected to bind to the target molecule related to the disease and influence its activity in a therapeutically beneficial manner. During this stage, High-Throughput Virtual Screening (HTVS) plays a crucial role in Lead Identification, leveraging High-Performance Computing (HPC) capabilities to efficiently screen large-scale database. HTVS employs two primary methods: Ligand-based and Structure-based virtual screening. Ligand-based Virtual Screening (LBVS) harnesses information from known ligands (molecules capable of binding to target proteins) to predict the activity of novel compounds. In contrast, Structure-based Virtual Screening (SBVS) relies on the target protein's structural information to identify potential ligands. Utilizing the target's 3D structure, SBVS employs molecular docking to sample the conformations of the complex, and scoring functions to assess their binding affinity. This affinity reflects the interaction strength between the molecules, with a higher affinity indicating a stronger interaction.

The primary contribution of this thesis centers on the scoring function, which is essential for the effectiveness and success of HTVS. Inaccuracies in the scoring function can compromise the entire HTVS process, while inefficiencies may hinder evaluating sufficient ligands within the time budget. Therefore, both accuracy and efficiency are of paramount importance for a scoring function. Currently, various scoring functions are available, such as physics-based, empirical, and knowledge-based scoring functions. However, most existing Scoring Functions compete with each other in prediction accuracy, while ignoring the importance of computational performance. Therefore, our work intends to introduce a new scoring function that considers both prediction accuracy and computational performance. This is particularly challenging as accuracy and computational performance often stand in opposition at the algorithm level, necessitating an optimal balance between them. Additionally, maximizing the utilization of available hardware and infrastructure within budgetary constraints is also a key consideration. To this end, our new scoring function addresses these aspects both at the algorithmic level and through the utilization of advanced hardware capabilities, particularly GPUs. Our primary objective is to achieve satisfactory accuracy while significantly enhancing computational performance, thereby enabling the rapid screening of extensive compound libraries. The scoring function we introduce, named DrugXGBScore, is a hybrid that merges the advantages of both knowledge-based and machine-learning scoring functions. More specifically, to achieve our objective, we integrated the optimized Drugscore2018, a Knowledge-based scoring function [2], with the recently popular machine learning algorithm eXtreme Gradient Boosting (XGBoost) [1]. Drugscore2018's simple structure allows for easy and efficient integration into our High-Performance Computing pipeline, while XGBoost contributes to further enhancing the prediction accuracy. To further enhance the computational efficiency, we custom-designed an HPC pipeline specifically for DrugXGBScore, harnessing the power of advanced Nvidia A-100 high-performance GPU to achieve our HPC objectives.

The predictive accuracy of the DrugXGBScore was assessed using the CASF-2016 dataset [3], normally used for scoring function evaluations. For computing performance, we conducted a screen test on a large dataset to compare the total running time and throughput between our HPC pipeline and a CPU-only setup. Finally, our comprehensive evaluation demonstrates that DrugXGBScore attains a mid-to-upper tier accuracy compared to other scoring functions in the CASF-2016 Power tests. In terms of computational performance, it remarkably screens all 28,500 decoys of a protein with approximately 8,000 atoms in just 8 seconds. For comparison, performing the same task using only a CPU would require nearly 27 hours, highlighting the significant efficiency of our HPC pipeline.

2. State of the Art

In academic, Scoring Functions are commonly classified into four categories: Physics-Based (e.g., GoldScore), Empirical (e.g., X-Score, AutoDock Vina), Knowledge-Based (e.g., DrugScore, DrugScore2018 [2]), and Machine Learning Scoring Functions (e.g., XGBoost [1]). Among these, the most relevant to our contributions are Knowledge-Based and Machine Learning Scoring Functions. Knowledge-based scoring functions extract pairwise potentials from the three-dimensional structures of numerous protein-ligand complexes, utilizing the inverse Boltzmann principle. They assume that the frequency of different atom pairs at specific distances indicates their interaction strength, which is translated into a distance-dependent potential of mean force (PMF). The final score can be derived from this PMF. Machine Learning Scoring Functions, on the other hand, employ machine learning algorithms, such as support vector machines, random forests, or gradient boosting, to predict the effectiveness of decoys as potential ligands directly.

The success of In-Silico Drug Discovery relies not only on the predictive accuracy provided by components such as Scoring Functions, but also on the computing performance offered by High-Performance Computing (HPC). In this context, we plan to use a hierarchical model to more effectively introduce HPC. This model systematically categorizes HPC into four distinct tiers, offering a logical and structured overview. At the broadest tier, the **Computer infrastructure level**, there's an emphasis on leveraging large-scale infrastructures, exemplified by supercomputers. Moving to the **Computer hardware level**, the spotlight is on exploiting parallel computing capabilities through technologies such as multi-core CPUs, GPUs, FPGAs, and specially designed hardware accelerators. At the **Computing framework level**, the focus is on incorporating parallel computing frameworks like MPI, OpenCL, and CUDA. Finally, at the **Computing algorithm level**, the essence is on devising high-performance algorithms tailored to specific computational tasks.

Notably, these levels are set to interplay and complement one another. For instance, when developing a high-performance algorithm for a specific task, it is essential to consider not only the task's execution, time complexity, and space complexity but also the effective use of the computing framework. This includes invoking tools like OpenMP for multithreading and CUDA for GPU acceleration, ensuring seamless alignment with the computing hardware and even underlying infrastructure, such as supercomputer, to maximize computational potential.

3. Hybrid Scoring Function - DrugXGBScore

The contributions of this Thesis are twofold. First, we propose DrugXGBScore, a Hybrid Scoring Function that linearly combines Knowledge-based and Machine Learning Scoring Functions. In this aspect, our focus is mainly on the computing algorithm level. While achieving satisfactory prediction accuracy, we aim to maximize computational efficiency by using simpler or High-Performance Computing (HPC)-friendly algorithms. Second, we integrate DrugXGBScore into our custom-designed HPC pipeline. This part concentrates on the computing framework and hardware level, employing appropriate computing frameworks, such as CUDA, to ensure DrugXGBScore works seamlessly with our HPC hardware, thus fulfilling our HPC objectives.

3.1. Overview of DrugXGBScore

Our proposed scoring function, DrugXGBScore, is a hybrid solution that integrates two distinct types of scoring equations. To achieve an optimal balance between prediction accuracy and computing performance, we combined Optimized DrugScore2018 [2], a Knowledge-based scoring function, with XGBoost [1], a Machine Learning based approach. The simpler structure of the Knowledge-based scoring function readily meets our HPC demands, while a well-trained Machine learning scoring function can further enhance prediction accuracy. Additionally, the XGBoost we selected features built-in high-performance computing capabilities [1].

The general workflow of the DrugXGBScore is shown in Figure 1. This process takes as input the 3D coordinates of all atoms in the Protein and Ligand, along with their SYBYL atom types. The final output is a score quantifying the Ligand's binding affinity to the Protein at a specific pose. A higher score signifies greater binding affinity. The process is comprised of four steps. In the first step, new conformations of the complex are calculated using a local optimization step. The second step involves scoring each new conformation with our Optimized DrugScore2018, selecting the highest score and its associated Feature Matrix. The third step utilizes the optimal Feature Matrix from the previous step as input for XGBoost, generating an XGBoost score. Finally, the fourth step combines the scores from Optimized DrugScore2018 and XGBoost linearly using a pre-trained parameter α to derive the final score.

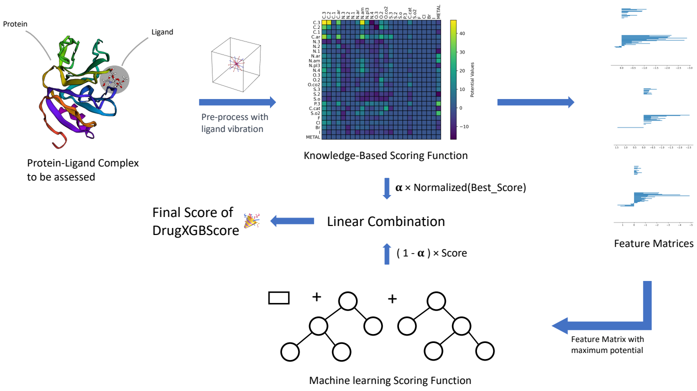
Protein
Ligand
Protein-Ligand Complex to be assessed
Pre-process with ligand vibration
Knowledge-Based Scoring Function
Potential Values
Feature Matrices
α × Normalized(Best_Score)
Final Score of DrugXGBScore
Linear Combination
(1 - α) × Score
+
+
Feature Matrix with maximum potential
Machine learning Scoring Function

Figure 1: The general workflow of DrugXGBScore

In this process, the Optimized DrugScore2018 we employed follows a training and inference process similar to the original DrugScore2018 [2]. The key innovations include parameter retuning and the introduction of a novel local optimization approach, termed the 'Ligand Vibration technique'. This technique primarily serves to improve dataset quality and address atom position uncertainty by adjusting the Ligand's position (moving in 27 directions from its original position) during the inference phase. Another highlighted aspect of this contribution is the features used for XGBoost input. We selected the feature matrix generated during the DrugScore inference process, which contains potential values of protein-ligand atom pairs extracted from the DrugScore model at various hit distances. Lastly, the parameter α used in the linear combination is determined through Bayesian search.

3.2. Deploying DrugXGBScore on the HPC Pipeline

To boost large-scale drug discovery, we integrated DrugXGBScore into the High-Performance Computing (HPC) pipeline, thereby enhancing High-Performance Virtual Screening (HPVS). This HPC pipeline is a specialized computational framework designed to handle complex and computing-intensive tasks. It employs advanced computing technologies and hardware, such as parallel computing and high-performance GPUs, to significantly accelerate our overall computing processes. During deployment, we strategically utilized our hardware for optimal parallel computing across both CPU and GPU platforms. We activated all CPU cores to operate independently for maximum efficiency. Concurrently, on the GPU, we tailored our algorithms for optimal execution using CUDA, aligning them with hardware capabilities. These approaches ensure superior computational performance and throughput. In practice, the input of our HPC pipeline consists of one or more proteins and their corresponding decoys that need to be screened. The output is the predicted score for each decoy, with higher scores indicating a greater binding affinity to the respective protein.

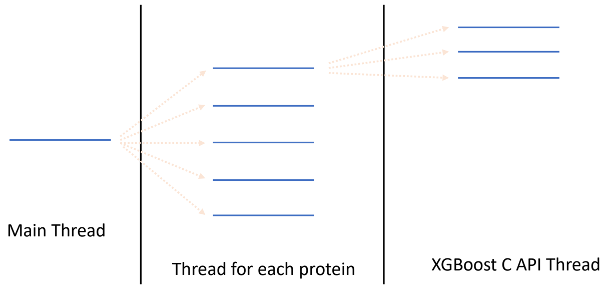

Figure 2: CPU Multiple Threads

Our CPU multi-threading strategy is depicted in Figure 2. The primary thread assigns distinct threads to simultaneously process each protein during virtual screening, free from data dependencies. In cases where the task involves only one protein, it can still be distributed across multiple threads to enable the concurrent screening of various ligands. For our Multi-threading Computing Framework, we utilized OpenMP, a well-established and user-friendly library with stable performance. Leveraging OpenMP, we launched 18 threads from the main thread on our device, which has 20 CPU cores. This approach was strategically chosen to prevent oversubscription and the resultant thread competition, as XGBoost's autonomous thread management could lead to more than 20 threads competing for CPU time, causing inefficiencies due to excessive context switching.

For GPU acceleration, we employed our cutting-edge GPU, the Nvidia A100 SXM4 40GB. Figure 3 visualizes the entire GPU accelerating process, using a single thread and protein as a representative example. This process is consistent across all other threads. In this figure, 'host' denotes the CPU and its memory, tasked with executing the main program, handling memory transfers, and launching the CUDA kernels. 'Device' refers to the GPU and its memory, which are dedicated to running the kernels. Our optimized DrugScore2018 model and proteins are initially transferred to GPU memory, followed by the ligands awaiting screening. The kernel is then invoked to compute the DrugScore and generate its associated feature matrix. After computation, the data are transferred back to the host memory, where the feature matrix is passed to the XGBoost C API. Finally, the resulting scores are linearly combined with the normalized DrugScore results. To maximize efficiency, we endeavored to perform all computing-intensive tasks within the CUDA kernel while minimizing memory transfer between host and device. Furthermore, as depicted by the dotted line in the center of the figure, we allocated a CUDA stream to each thread, facilitating concurrent execution of CUDA kernels and data transfers.

4. Experimental Results

Our objective is to maximize computing performance while maintaining satisfactory prediction accuracy. Therefore, our experimental tests are centered around these two critical aspects. Firstly, to assess prediction accuracy, we employ CASF-2016 Power test [3] as a benchmark, comparing DrugXGBScore's accuracy with other Scoring Functions. Then, to evaluate computing performance, we screen a large test set using our HPC pipeline, contrasting the total running time and throughput with a CPU-only ap-

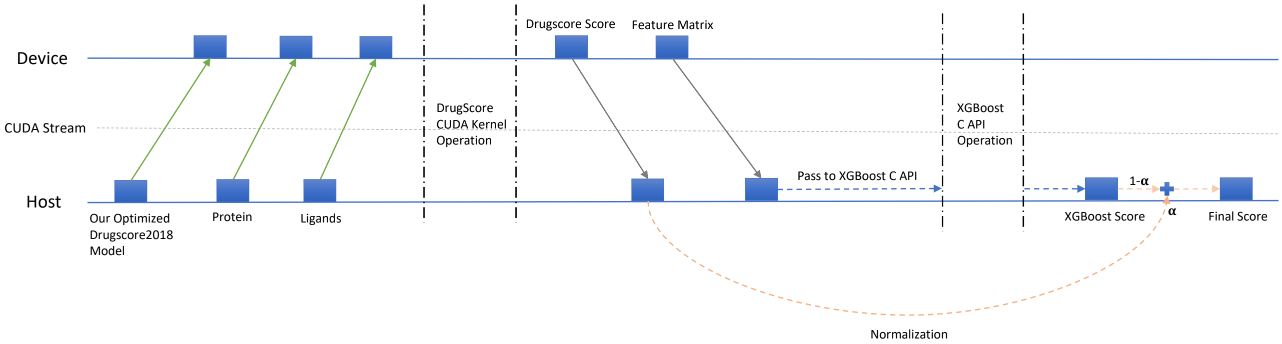

Figure 3: GPU Accelerating Process - Single CUDA Stream

proach.

Prediction Accuracy Test. CASF-2016 includes four power tests to evaluate scoring functions:

- **Scoring Power**: Assesses scoring function proficiency in producing binding scores with a linear correlation to experimental values of protein-ligand complexes, using the Pearson correlation coefficient (R) as the primary metric.
- **Ranking Power**: Measures the ability to accurately order known ligands of a target protein based on binding affinities, given exact binding poses, with Spearman's rank correlation coefficient (ρ) as the primary metric.
- **Docking Power**: Evaluates the capability of distinguishing native ligand binding poses from computer-generated decoys, using success rates of binding poses as the metric.
- **Screening Power**: Tests the proficiency in identifying true binders among random ligands with varied poses, using the Average enrichment factor as the primary metric.

Table 1 presents the CASF-2016 Power Test results, comparing the prediction accuracy of Optimized DrugScore2018, XGBoost, and the final DrugXGBScore. The metrics used are the Pearson correlation coefficient (R) for Scoring Power, Spearman correlation coefficient (ρ) for Ranking Power, success rate of the top1 binding pose for Docking Power, and average enrichment factor among top 1% for Screening Power. This result indicates that our Optimized DrugScore2018 excels in Docking and Screening Power, while XGBoost outperforms in numerical predictions, as evidenced by its effective Scoring Power. Significantly, the most critical metric is the average enrichment factor of Screening Power, as the primary function of DrugXGBScore involves screening, namely identifying optimal ligands from a large-scale database for in-vitro testing. The results demonstrate that linearly combining Optimized DrugScore2018 with XGBoost into DrugXGBScore markedly improved overall screening power. This approach resulted in a notable increase in the average enrichment factor, reaching 4.52, compared to using either Optimized DrugScore2018 or XGBoost alone.

Moreover, based on the data presented in the available thesis (with all other methods detailed in the full documents), our DrugXGBScore achieves near-optimal performance in Scoring and Ranking Power and ranks in the upper-middle tier for Docking and Screening Power. We believe these results align well with the prediction accuracy requirements for this drug discovery task.

Computing Performance Test. Subsequently, we conducted a large-scale computational performance test of our HPC pipeline, comparing its total running time and throughput with a CPU-only setup to evaluate the performance improvements achieved. This was followed by two sets of experiments designed to evaluate DrugXGBScore's performance across different scenarios.

- In the first set, we utilized 57 proteins for screening, along with their corresponding 1,624,500 ligands.
- For the second set, our focus was on screening potential decoys for a single protein, ACETYLCHOLINESTERASE, which has 8313 atoms, is identified by PDB Code 1E66, and includes 28,500 decoys.

The results are displayed in Figure 4. Panel A shows the total running time for both test sets,

Metrics	Optimized DrugScore2018	XGBoost	DrugXGBScore
Scoring Power (R)	0.604	0.733	0.714
Ranking Power (ρ)	0.618	0.600	0.625
Docking Power (Success Rate Top1)	82.5%	43.2%	75.1%
Screening Power (Top 1%)	3.08	2.35	4.52

Table 1: CASF-2016 Power Test Results

with blue indicating CPU-Only usage and red representing GPU acceleration. Panel B illustrates the throughput. Given the substantial differences in both running time and throughput, we use a Logarithmic Scale for the Y-axis to facilitate visual comparison. On the CPU-only setup, screening 28,500 decoys of a single protein with approximately 8,000 atoms requires 105,023 seconds, about 27 hours. For all 57 test proteins with their 1,624,500 decoys, the process takes nearly 55 days. However, the efficiency significantly improves with our GPU acceleration. Screening 28,500 decoys for a single protein now takes just 8.51 seconds, while completing the entire Screening Power test requires only 184 seconds. The throughput also astonishingly reaches approximately 3,300 and 8,700 ligands per second, respectively. Overall, our HPC pipeline's performance exhibited nearly four orders of magnitude increase compared to a CPU-only setup. It is noteworthy that while the throughput with CPU-only shows little variation between different scenarios, it differs notably post-GPU acceleration. This discrepancy is attributed not only to the significant variation in the number of atoms among different proteins but also to the startup time involved in GPU acceleration. Such non-computational overhead includes tasks like memory allocation, freeing up GPU memory, and memory transfer.

5. Conclusions

To address the challenges of prolonged timelines, significant costs, and high failure rates in drug discovery, we introduced a hybrid scoring function, DrugXGBScore, for high-performance virtual screening. This scoring function identifies potential drug candidates for the target protein from a large-scale drug molecule database, which are then forwarded for experimental testing in a wet laboratory.

In this process, our primary goal is to maintain satisfactory prediction accuracy while maximizing computing performance. Correspondingly, the experimental results show that DrugXGBScore ranks in the upper-middle tier among other scoring functions in CASF-2016 in terms of prediction accuracy. On the other hand, regarding computing performance, our HPC pipeline achieved an overall enhancement of four orders of magnitude compared to a CPU-only setup, which is a significant achievement.

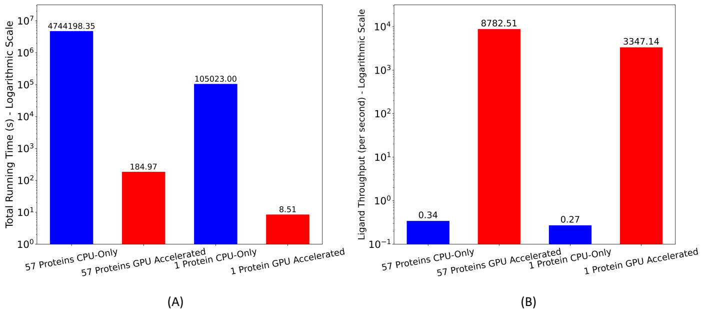

Figure 4: Performance Comparison: Our HPC Pipeline vs CPU-Only Setup

References

[1] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794.

[2] Jonas Dittrich, Denis Schmidt, Christopher Pfleger, and Holger Gohlke. Converging a knowledge-based scoring function: DrugScore ^{2018}. 59(1):509–521.

[3] Minyi Su, Qifan Yang, Yu Du, Guoqin Feng, Zhihai Liu, Yan Li, and Renxiao Wang. Comparative assessment of scoring functions: The CASF-2016 update. 59(2):895–913. Publisher: American Chemical Society.